

CURB: Civic Urban Reasoning Benchmark

Oliver R. Fox and Ellis Solaiman

Newcastle University

1 Science Square

Newcastle upon Tyne, NE4 5TG, UK

{o.fox3, ellis.solaiman}@newcastle.ac.uk

Abstract

Civic Urban Reasoning Benchmark (CURB) evaluates fine-grained evidence-to-claim reasoning over public and official civic information sources. The task asks whether a short civic evidence item supports, refutes, or leaves insufficient evidence for a claim. Civic claims are hard to verify because near-identical wording can hide a small change in location, time, status, or cause that flips the truth of the claim, so high lexical overlap is not enough; **CURB** systematically targets these distinctions, including local entities, status changes, temporal constraints, causes, and unsupported extra detail. The current split contains 30 evidence items and 90 balanced evidence–claim pairs across transport, roadworks, weather, and crime. We compare neural baselines with **Logical, Structural and Semantic text Interpretation (LaSSI)**, a logic-grounded explainable pipeline using ontology-backed typing, structural rewriting, and ex-post directional logical comparison. **LaSSI** reaches 85.6% accuracy and 0.858 Macro-F1, narrowly ahead of strong **Natural Language Inference (NLI)** cross-encoder baselines (up to 83.3%), with its clearest gains on location, time, and cause distinctions. The results also show clear limits: paraphrase and status changes remain competitive or stronger for neural models, and telegraphic roadworks notices stress the parser to the point that a cross-encoder overtakes **LaSSI** on that domain.

1 Introduction

Civic claim verification is a concrete and consequential problem: residents, services, and automated assistants increasingly need to check short public-interest claims, such as whether a road is closed, a station entrance is shut, or a crime case is resolved, against public and official civic information feeds. These claims are difficult to verify precisely because they are often highly similar in wording yet truth-sensitive to small details: a change in location, time, status, or cause can flip

a claim from true to false while leaving most of the sentence unchanged. This makes civic verification a sharper test than general misinformation detection (Battista et al., 2025), and one for which automated fact-checking and **Natural Language Inference (NLI)** pipelines must establish strict, directional support or contradiction between retrieved evidence and a target claim (Guo et al., 2022).

Two short examples make this concrete. A piece of civic evidence stating that a road is closed “near Edward Place” supports a claim that it is closed “on or near Edward Place,” but the reverse entailment does not necessarily hold. Similarly, an event at “Parking Area, Newcastle” and an event at “Central Station, Newcastle” share city-level context but denote completely distinct granular locations.

Contemporary approaches often use pre-trained sentence transformer models to compute dense vector embeddings and then compare premise–claim pairs with cosine similarity or related scoring functions (Reimers and Gurevych, 2019). These systems provide strong semantic generalisation and are likely to outperform symbolic pipelines on many fluent paraphrases and broad lexical variations. However, embedding similarity is not by itself a guarantee of directional entailment: two civic statements may share most of their vocabulary while differing in a location, time, causal explanation, or lifecycle status that changes the truth condition of the claim (Section 6).

Civic Urban Reasoning Benchmark (CURB) is not designed to show that logic-based reasoning beats neural **NLI** everywhere; it is designed to expose where civic evidence-to-claim reasoning depends on exact local entities, temporal constraints, causes, statuses, and unsupported details.

The main contributions of this work are as follows:

- The **CURB** dataset, a civic evidence-to-claim reasoning benchmark designed to test whether

public and official civic evidence strictly supports, refutes, or provides insufficient evidence for a fine-grained claim, with claims systematically varied along location, time, cause, status, negation, paraphrase, and unsupported extra details.

- An evaluation of sentence transformers, cross-encoders, and [Large Language Models \(LLMs\)](#) on [CURB](#), which show that dense bi-encoders largely fail to register contradiction under a fixed decision rule, while [NLI](#) cross-encoders are strong and efficient baselines that nonetheless still trail on fine-grained location and time distinctions.
- An evaluation of [Logical, Structural and Semantic text Interpretation \(LaSSI\)](#) (detailed in Section 4), a hybrid explainable logical rewriting method grounded in the [Hub Ontology for Knowledge \(HONK\)](#), showing that explicit logical structure can reduce false-positive support judgements in fine-grained civic cases when parsing and grounding succeed, while remaining brittle on telegraphic notices such as roadworks.

2 Related Work

In this work, civic evidence is drawn from public and official information sources, such as government, transport, weather, and policing agencies. We frame verification as textual entailment, where one fragment of text logically implies another ([Korman et al., 2018](#)). In computational linguistics, [NLI](#) determines whether a given hypothesis can be logically inferred from a premise. This is typically categorised as entailment, contradiction, or neutral ([Dagan et al., 2005](#)).

Fact-Checking and [NLI](#) Benchmarks Existing benchmarks have extensively explored automated fact-checking and [NLI](#) ([Sathe et al., 2020](#); [Bowman et al., 2015](#)). These datasets primarily focus on broad-domain general knowledge, generic news articles, or encyclopaedic evidence. While these resources have driven significant advances in [Natural Language Processing \(NLP\)](#), they do not isolate the fine-grained civic truth conditions that turn on *local entities, times, causes, statuses, and unsupported details*: a small change can flip a claim’s truth while leaving most of its wording intact. [CURB](#) is built to target this gap.

Weaknesses in Dense Embeddings Recent advancements in [LLMs](#) and dense vector embeddings have shown impressive generative and semantic matching capabilities. These models form strong baselines for [CURB](#), especially when a claim is a fluent paraphrase of the evidence or requires broad lexical generalisation. Their limitation is that similarity-oriented scoring can conflate lexical overlap with entailment in cases involving entity swaps, negation, asymmetric spatial phrasing, or unsupported extra detail ([Fox et al., 2025](#); [Song et al., 2026](#)).

Knowledge Graphs and Symbolic Reasoning Neuro-symbolic approaches integrate [Knowledge Graphs \(KGs\)](#), typed entities, or symbolic reasoning into language-processing pipelines ([Chaudhuri et al., 2021](#); [De Giorgis et al., 2025](#)). [KGs](#) can support systems such as [LaSSI](#) and [LLMs](#) by providing structured background knowledge ([Zhang et al., 2025](#)). Their appeal in civic verification is transparency: a system can expose which entities, relations, or logical atoms contributed to a support or contradiction judgement. Their weakness is also well known: they depend on parser quality, ontology coverage, and hand-engineered or curated mappings between language and formal structure.

3 CURB Task and Dataset

3.1 Task Definition

Given a short civic evidence item E from a public or official civic source and a manually authored claim C , the task is to predict one of three labels: *Supported*, *Refuted*, or *Not Enough Evidence*. The direction is fixed as $E \Rightarrow C$: the benchmark asks whether the evidence licenses the claim, not whether the two texts are semantically similar in both directions.

Unlike full automated fact-checking pipelines, which must also retrieve or select evidence ([Guo et al., 2022](#)), [CURB](#) fixes the evidence item in advance and focuses on the reasoning step: whether that evidence supports, refutes, or leaves insufficient evidence for the claim.

Label Semantics *Supported* means that the claim is licensed by the evidence without adding material information. *Refuted* means that the claim contradicts an explicit fact in the evidence, such as a different location, status, time, or cause. *Not Enough Evidence* means that the claim is plausible but contains a detail that the evidence does not

establish. These labels follow the broad NLI convention, but are specialised here for single-source civic verification.

3.2 Data Construction

The **North East Evidence Tracker (NEET)** dataset was curated using an automated pipeline that harvests civic evidence from the North East of England, specifically targeting Newcastle upon Tyne and the Tyne and Wear area; the dataset and evaluation outputs are publicly available¹ alongside the codebase.² We focus on four domains relevant to smart-city information services (Battista et al., 2025): transport, roadworks, weather, and crime. This is achieved by scraping four heterogeneous public data sources:

- **Nexus Metro Updates:** service disruption and travel-update notices.
- **Newcastle City Council Roadworks:** active and planned roadworks, restrictions, diversions, and expected end dates.
- **Met Office:** local forecast strings and weather metrics.
- **Police.uk:** reported crime and anti-social behaviour records.

Each record stores the raw evidence snippet, its source name and URL, the scrape timestamp, an extracted focal location where available, the authored claim, the gold label, and the modification strategy; the full per-row schema is given in Appendix A.1.

3.3 Claim Generation and Modification Strategies

To establish our testing paradigm, exactly three claims were generated for each of the 30 evidence snippets, corresponding directly to the three standard NLI classification labels: Supported (Entailment), Refuted (Contradiction), and Not Enough Evidence (Neutral). The current split therefore contains 30 evidence items and 90 evidence–claim pairs, with one Supported, one Refuted, and one Not Enough Evidence claim per evidence item. The domain distribution is transport (8 evidence items), weather (8), crime (7), and roadworks (7). Each claim is produced by one of seven modification strategies: paraphrase, negation, time change, location change, cause change, status/outcome

change, and unsupported extra detail (defined in Appendix A.2).

Annotation and Validation All 90 claims were authored by the first author following the modification-strategy protocol above; each strategy determines its target label by construction (e.g. negation yields *Refuted*, an added unverified detail yields *Not Enough Evidence*). As validation, the second author independently reviewed every evidence–claim–label triple, checking both the label and that the claim realised its strategy; disagreements were resolved by discussion and the affected claims revised. No external annotators were employed (Section 8).

4 Logic-Based Verification Method

We evaluate **LaSSI** (Fox et al., 2025) as an interpretable, logic-grounded method for **CURB**, not as a replacement for contemporary neural NLI. Its accuracy depends on successful parsing, chunking, and ontology grounding. The pipeline uses an ontology-backed layer (**HOnK**) that aggregates lexical, commonsense, spatial, paraphrase, and lifecycle resources; we describe only the functionality needed for this paper.

4.1 HOnK (Ontological Grounding)

Contemporary NLP models may encode broad geographic and world knowledge, but they do not necessarily provide stable, inspectable grounding for local civic entities without task-specific data or retrieval (Lavrinovics et al., 2025).

HOnK aggregates and clusters concepts from multiple knowledge resources, including ConceptNet (Speer et al., 2017), GeoNames (Li et al., 2025), WordNet (Miller, 1995), Wiktionary (Foundation, 2026), and project-specific civic vocabularies. Within **LaSSI**, **HOnK** is loaded as a pre-built ontology resource and queried through in-memory adjacency structures for synonymy, supertypes, subtypes, paraphrases, lifecycle states, and spatial relations.

By grounding these civic entities, the pipeline’s subsequent logical reasoning phase can evaluate semantic equivalence through structural identity and spatial constraints. For example, in our ‘Location Change’ modification strategy, **HOnK** prevents the system from confusing two nearby but distinct streets, enabling **LaSSI** to correctly flag the claim as *Refuted*, whereas dense vector models typically yield a high false-positive similarity due to

¹<https://osf.io/fu4bg/>

²<https://github.com/0lliefox/LaSSI/tree/v4.3>

shared vocabulary contexts (Table 2, where **LaSSI** reaches 100.0% on location mismatch against 0.0% for the strongest bi-encoder).

4.2 LaSSI (Logic Rewriting and Verification)

LaSSI verifies each evidence–claim pair in four conceptual steps. First, the terse civic text is *parsed and chunked*: a row-chunk pre-parser protects atomic spans (ISO timestamps, date ranges, percentages, decimals, and unit-bearing measurements) and assigns chunk roles before dependency parsing with StanfordNLP (Qi et al., 2019), in the spirit of cascaded finite-state and clinical information-extraction systems for semi-structured text (Hobbs et al., 1997; Savova et al., 2010). The resulting tokens are normalised into a typed **Multi-Word Entity Unit DataBase (meuDB)** that bounds compound names such as “Newcastle city centre” as one unit. The detailed chunk-role and rewrite machinery is given in Appendix A.

Second, these typed entities are *grounded* through **HOnK**, and third, *rewritten into logical atoms*: an extended **First-Order Logic (FOL)** representation in which each sentence is a set of typed atoms carrying relational, spatial, temporal, causal, and status properties. We formalise the evidence as $E = \{e_1, \dots, e_n\}$ and the claim as $C = \{c_1, \dots, c_m\}$. Spatial phrasing is typed at this stage; for instance “on or near” a place is represented as a disjunctive *at/near* location, which is what lets the comparison preserve the asymmetric spatial entailment discussed in Section 7 (mechanics in Appendix A).

Fourth, the *ex post* phase *compares the atoms directionally*, using **HOnK/TBox** expansion together with paraphrase and lifecycle resources under closed-world-style scoring; the model-counting procedure is detailed in Appendix A.7. The comparison is *directional*: $E \Rightarrow C$ holds only when the truth conditions of every atom in C are satisfied by E . This lets **LaSSI** distinguish a claim that is strictly weaker than the evidence (e.g. evidence asserts an exact location, claim asserts “on or near” it) from one that asserts more than the evidence supports (e.g. the evidence is silent on a property the claim specifies), a distinction that symmetric similarity scores cannot represent.

LaSSI maps this directional comparison into the benchmark label space as follows: **Supported** when all logical atoms in C are satisfied by E ; **Refuted** when an explicit logical or categorical contradiction is found (e.g. a spatial mismatch); and

Not Enough Evidence when only a partial graph intersection holds, proving neither entailment nor contradiction.

This design makes **LaSSI** interpretable: when it succeeds, the supporting or conflicting atoms can be inspected. Its central limitation is that malformed dependency parses or missing ontology coverage can produce malformed logical atoms, causing the later verification stage to under-support true paraphrases, miss contradictions, or return partial scores. This limitation is especially relevant for terse civic notices that are not well-formed prose.

5 Experimental Setup

5.1 Baselines

To evaluate **LaSSI** against contemporary neural approaches, we compare it with three baseline paradigms.

Sentence Transformers (Bi-Encoders) Bi-encoder sentence transformers encode the evidence and claim independently, then compare the two embeddings with cosine similarity or a related score (Reimers and Gurevych, 2019). They are efficient and suitable for retrieval-style semantic matching, but the score is symmetric and does not directly encode the direction $E \Rightarrow C$. We evaluate all-MiniLM-L6-v2 and all-MiniLM-L12-v2 (Wang et al., 2020), and all-mpnet-base-v2 (Song et al., 2020).

NLI Models (Cross-Encoders) Cross-encoders process the evidence and claim jointly, allowing token-level interactions across the pair. They are included as neural baselines for pairwise sentence classification and reranking. For the three-way **NLI** models, the predicted class (*entailment / neutral / contradiction*) is read directly from each model’s own label set and mapped *a priori* to *Supported / Not Enough Evidence / Refuted* respectively. This correspondence is fixed in advance and is not tuned on any portion of **CURB**, so it neither inflates nor handicaps these baselines. We evaluate bge-reranker-v2-m3 (Chen et al., 2024), nli-deberta-v3-base (Reimers and Gurevych, 2019; HuggingFace, 2025a), and nli-MiniLM2-L6-H768 (Wang et al., 2020; HuggingFace, 2025b).

LLMs (Zero-Shot) We use structured-rationale prompting: each **LLM** is prompted to produce a brief justification and then a label/score, returned as a JSON object. This evaluates whether

modern LLMs can approximate the CURB label space zero-shot, without task-specific fine-tuning. Via a local Ollama instance (Ollama, 2026) we evaluate gemma4:e2b (Manik and Wang, 2026), llama3.2:3b (Grattafiori et al., 2024), and qwen2.5:7b (Yang et al., 2024).

5.2 Thresholds and Label Mapping

LaSSI produces a directional logical comparison matrix. Cell $[i][j]$ is interpreted as the directional support value from S_i to S_j , so the first row asks how much the source evidence S_0 licenses each claim sentence. For discrete matrix-style outputs (LaSSI and the LLM matrices, but not the threshold-mapped encoders), the label derives from both directions of the computed matrix: 0.0 in either direction marks contradiction, 1.0 in either direction full support, and otherwise insufficient evidence. When both directions are partial, the benchmark’s curated directional expectations resolve the pre-specified asymmetric cases, such as “on or near X” supporting “near X”, as *Supported*, since a three-way label space cannot itself separate asymmetric support from insufficient evidence. This resolution is a fixed *a priori* rule: the asymmetric cases are part of the benchmark definition, specified before any system was evaluated, and the rule is applied identically to every matrix-style system rather than being tuned per model (Section 8).

As cosine similarity measures symmetric semantic proximity rather than asymmetric entailment (Reimers and Gurevych, 2019; Fox et al., 2025), there is no universal threshold for this task. We therefore adopt a fixed *a priori* decision rule for sentence-transformer outputs rather than tuning on data: scores $\geq \tau_{\text{high}} = 0.8$ are mapped to *Supported*, scores $\leq \tau_{\text{low}} = 0.2$ to *Refuted*, and intermediate scores to *Not Enough Evidence*. These cut-offs are stated up front and are not selected against any portion of CURB, which avoids the appearance of fitting the baseline to the test answers; the trade-off is that a symmetric, non-directional signal is unlikely to be squeezed for additional accuracy without leaking information from the evaluation set. Cross-encoders and LLMs output discrete labels directly via logits or prompted JSON, so no thresholding is applied to them. A *post-hoc oracle* analysis over all threshold pairs (Appendix B) shows that thresholds selected with full knowledge of the test labels lift the best bi-encoder only to 64.4%, so the deficit is representational. A validation-split threshold sweep on a larger CURB release is left

as future work (Section 8).

5.3 Evaluation Metrics

Evaluation is reported using overall accuracy and Macro-F1, with additional slices by civic domain and modification strategy. Domain-level slices are needed because weather, roadworks, transport, and crime differ substantially in surface form. Modification-strategy slices are needed because the central hypothesis concerns specific reasoning phenomena, not only headline accuracy. Due to the small benchmark, each table reports the number of pairs n underlying every slice, and percentage differences should be read against these denominators.

Environment All experiments were run on an Apple M4 Pro machine with 48 GB of unified memory. Sentence transformers and cross-encoders were run through the sentence-transformers/PyTorch stack, and the zero-shot LLMs through a local Ollama instance with Metal acceleration. Reported times are per evidence–claim comparison, averaged over the benchmark. They are not like-for-like across paradigms: the LaSSI figure includes dependency parsing, ontology grounding, and logical rewriting, whereas the encoder figures reflect single forward passes over the sentence pair.

6 Results

Quantitative Results Table 1 reports the main comparison over 30 civic evidence items and 90 evidence–claim pairs. LaSSI obtains the highest overall score, with 85.6% accuracy and 0.858 Macro-F1. The margin over the best neural baselines is, however, narrow; 2.3 points over nli-MiniLM2-L6-H768 corresponds to just two additional correct predictions on this split, so we do not claim statistical significance for the ordering (Section 8): the NLI cross-encoders nli-MiniLM2-L6-H768 (83.3%, 0.832) and nli-deberta-v3-base (81.1%, 0.811) are close behind, and the strongest LLM, gemma4:e2b, reaches 71.1% (0.692). The dense bi-encoders are the weakest paradigm here (44.4–47.8%), as a single symmetric cosine score cannot encode the directional decision $E \Rightarrow C$. We therefore read LaSSI’s lead as a *narrow, structure-dependent, in-domain* advantage rather than a general one: as Section 8 details, its rules and vocabulary were refined using

Table 1: Overall performance across the **CURB** test set ($n = 90$ evidence–claim pairs; one percentage point ≈ 0.9 pairs). For accuracy and Macro-F1, highest values are blue and lowest are red; for time, fastest is blue and slowest is red.

Paradigm	Model	Correct	Accuracy (%)	Macro-F1	Avg. Time (s)
Sentence Transformers	all-MiniLM-L6-v2	41/90	45.6	0.369	0.169
	all-MiniLM-L12-v2	40/90	44.4	0.362	0.247
	all-mpnet-base-v2	43/90	47.8	0.382	0.240
Cross-Encoders	bge-reranker-v2-m3	51/90	56.7	0.470	0.261
	nli-deberta-v3-base	73/90	81.1	0.811	0.269
	nli-MiniLM2-L6-H768	75/90	83.3	0.832	0.130
Generative LLMs	gemma4:e2b	64/90	71.1	0.692	17.625
	llama3.2:3b	57/90	63.3	0.640	3.206
	qwen2.5:7b	59/90	65.6	0.569	6.314
LaSSI		77/90	85.6	0.858	6.309

Table 2: Accuracy (%) sliced by modification strategy for representative models, with the raw number correct beneath each percentage and the number of pairs n per slice. The smallest slices (cause mismatch, status change) shift by 16.7 and 25 points respectively per additional correct prediction. Highest per column in blue; lowest in red.

Model	Paraphrase ($n=30$)	Negation ($n=14$)	Location Mismatch ($n=10$)	Time Mismatch ($n=10$)	Cause Mismatch ($n=6$)	Status Change ($n=4$)	Unsupported Extra Detail ($n=16$)
all-mpnet-base-v2	86.7 26/30	0.0 0/14	0.0 0/10	0.0 0/10	100.0 6/6	0.0 0/4	68.8 11/16
nli-deberta-v3-base	63.3 19/30	100.0 14/14	70.0 7/10	70.0 7/10	100.0 6/6	100.0 4/4	100.0 16/16
gemma4:e2b	90.0 27/30	100.0 14/14	80.0 8/10	30.0 3/10	33.3 2/6	75.0 3/4	43.8 7/16
LaSSI	73.3 22/30	85.7 12/14	100.0 10/10	90.0 9/10	100.0 6/6	75.0 3/4	93.8 15/16

CURB, so this figure is an optimistic in-domain development estimate.

The effect that drives this is visible in Table 2. **LaSSI** is strongest precisely on the slices that require local structure rather than broad semantic relatedness: location mismatch (100.0%), time mismatch (90.0%), and cause mismatch (100.0%, tied with the bi-encoder and nli-deberta). It remains competitive on unsupported extra detail (93.8%). It is weaker on paraphrase (73.3%), where gemma4:e2b (90.0%) and all-mpnet-base-v2 (86.7%) lead, and on status change (75.0%), where nli-deberta-v3-base reaches 100.0%. Notably, nli-deberta-v3-base matches or beats **LaSSI** on negation, status change, and unsupported extra detail (all 100.0%).

Table 3 shows that, on this split, **LaSSI** reaches 100.0% on weather and crime and remains strong on transport (83.3%), but is weakest on roadworks (57.1%). Roadworks is in fact the one domain where a neural baseline clearly overtakes it: nli-deberta-v3-base reaches 81.0%. We treat this as diagnostically important rather than as a failure to obscure (Section 8): roadworks notices are the most telegraphic civic form in **CURB**, and

they stress the dependency-parser-led rewriting that **LaSSI** depends on.

The label-level pattern (Figure C1, Appendix C) is also clear. Dense bi-encoders almost never predict *Refuted* on this split, even when the claim contradicts a local entity, time, status, or cause (the *Refuted* column for all-mpnet-base-v2 is empty); the **NLI** cross-encoders, by contrast, do recover contradiction. **LaSSI** makes a conservative trade-off: it almost never turns a refuted or insufficient-evidence case into a false *Supported* label, but it sometimes under-supports true supported claims by returning *Not Enough Evidence*.

Directional Matrix Interpretation The logical and **LLM** matrix outputs are interpreted as directional support values, where cell $[i][j]$ measures support from S_i to S_j . A first-row value of 0.5 is therefore not automatically a failed support judgement. In asymmetric cases, especially where one sentence is more specific than the other, the correct evidence-to-claim value can be partial while the pair is still labelled *Supported*. The evaluation therefore uses the curated directional matrix expectations for these discrete matrix-style outputs. Con-

Table 3: Accuracy (%) sliced by civic domain for representative models, with the raw number correct beneath each percentage and the number of pairs n per slice. Highest per column in blue; lowest in red.

Model	Transport ($n=24$)	Roadworks ($n=21$)	Weather ($n=24$)	Crime ($n=21$)
all-mpnet-base-v2	54.2 13/24	47.6 10/21	37.5 9/24	52.4 11/21
nli-deberta-v3-base	79.2 19/24	81.0 17/21	66.7 16/24	100.0 21/21
gemma4:e2b	70.8 17/24	71.4 15/21	66.7 16/24	76.2 16/21
LaSSI	83.3 20/24	57.1 12/21	100.0 24/24	100.0 21/21

tinuous sentence-transformer, reranker, and cross-encoder scores keep the fixed threshold mapping described in Section 5.

Error Analysis: Telegraphic Notices Roadworks notices contain terse action fragments, abbreviations, duplicated punctuation, dimensional measurements, date ranges, and traffic-management labels. Examples such as “*Demolish FW2 rebuild FW4. 15m x lowering*” and “*Excavate on existing valve chamber to renew cover.*” are meaningful to a civic reader but difficult for dependency parsing and row reconstruction. The roadworks results in Table 3 show that these failures remain visible despite protected-span chunking and structural rewrites.

Transport failures have a similar source. Long notice rows can repeat headings, merge date ranges with prose, and bury the causally relevant clause inside a run-on announcement. When the parser loses the boundary between header, time range, and action, later logical comparison cannot recover the intended atoms reliably. Weather rows are also compact, but their structure is more regular, and the current protected-span and weather-vocabulary handling is sufficient for this split.

7 Analysis and Discussion

Comparison with Neural Baselines The results support a narrow, not universal, advantage for LaSSI. Neural baselines remain strong where the task resembles fluent paraphrase or broad semantic matching: gemma4:e2b reaches 90.0% and all-mpnet-base-v2 86.7% on paraphrase in Table 2, compared with 73.3% for LaSSI. The NLI cross-encoders are especially competitive overall (nli-MiniLM2-L6-H768 83.3% and nli-deberta-v3-base 81.1%, against LaSSI’s 85.6% in Table 1), and they also handle several categorical status changes better than the current

logical pipeline.

The logic-grounded advantage appears where a small civic detail changes the truth condition. In Table 2, LaSSI leads on location mismatch, time mismatch, cause mismatch, and unsupported extra detail. This is the benchmark’s intended pressure point: high lexical overlap is compatible with contradiction or insufficient evidence, so a system must decide whether the evidence licenses the claim in the direction $E \Rightarrow C$.

Asymmetric Entailment Civic evidence often contains asymmetric phrasing. A more specific record can license a weaker claim, while the weaker claim may not license the full record in reverse. The directional support value from S_i to S_j is therefore asymmetric rather than a symmetric similarity score. Some supported cases correctly have a first-row partial value, such as 0.5, because the pair differs in specificity even though the evidence is sufficient for the benchmark claim. Treating every partial value as a failure would erase the very asymmetry that CURB is designed to test.

Granular Spatial Recognition CURB also tests local spatial identity. Mentions such as “*Parking Area, Newcastle*” and “*Central Station, Newcastle*” share broad city context but denote different local places. In Table 2, LaSSI reaches 100.0% on location mismatch, while all-mpnet-base-v2 reaches 0.0%. This is consistent with the intended role of HOnK-backed typing and spatial comparison: lexical overlap is not enough when the local entity changes.

Categorical Status and Lifecycle Changes Status changes remain mixed. The lifecycle resources let LaSSI represent mutually exclusive states such as under investigation, no suspect identified, charged, delayed, normal service, open, or closed. However, this only works when the rele-

vant phrase is covered and attached to the right predicate. Table 2 shows LaSSI at 75.0% on status change, level with gemma4:e2b and below nli-deberta-v3-base (100.0%). This is a useful counterexample to any broad symbolic-superiority claim.

Weather Forecast Structure Weather is a success case in the current split (100.0% in Table 3) despite also being compact and semi-structured. The reason is that a forecast row draws on a small, *closed* surface schema: a source/report header, a named condition, and a fixed inventory of metrics (temperature, wind speed, precipitation probability) attached to an ISO timestamp. The protected-span pre-tokenizer treats each timestamp, decimal, percentage, and unit-bearing measurement as one atomic token before splitting, the chunk profiler assigns the leading clause a report-header role and the metric fragments an attribute role, and the weather paraphrase and lifecycle resources align lexical variants and mutually exclusive states (for example “rain”/“precipitation” and wet/dry, cloudy/clear). As the vocabulary and field layout are enumerable, these components can pre-structure the row into clean atoms.

Roadworks and Transport Notice Structure Roadworks is harder because it is *open-ended* rather than enumerable: a roadworks body carries free imperative actions, domain abbreviations, and dimensioned operations for which no closed paraphrase or lifecycle vocabulary exists, so the protected-span trick cannot help. The difficulty is the open verbal and argument structure itself. The per-case behaviour is instructive. Among roadworks rows, only one (“Roadworks on Fern Drive: Abandon approx 11m of 6" SI and replace with approx 11m of LP PE...”) has its supported claim fully recovered: it carries an explicit coordinated action (*abandon ...and replace with ...*) with reasonably complete arguments, which the row-chunk pipeline’s action-object promotion can re-assemble into a single kernel. The remaining rows defeat reassembly: bare juxtaposed imperatives without a conjunction (“Demolish FW2 rebuild FW4”), opaque abbreviations, dimensioned operations (“15m x lowering”), purpose-clause chains (“Excavate on existing valve chamber to renew cover”), and double full stops all yield incomplete atoms, so the supported claim collapses to a partial score and the domain falls to 57.1%.

Transport shows the mirror image. Rows that

lead with a place-name header glued onto a prose body (for example “*Northumberland Park Car Park There is a temporary reduction ...*”) are correctly handled: the pipeline separates the header from the clause so the main predicate parses, and all six such rows are labelled correctly. The two failing rows instead lead with a *date* header and then repeat the closure clause several times in one field (“*Monday, 13 April 2026- Friday 17 April Five day line closure: Five day line closure ... No trains will run ...*”); here the boundary between header, time range, and action is lost, and the supported and refuted claims both collapse to an indeterminate 0.5. These are not failures of logical comparison alone: when the row-chunk pipeline or dependency parse produces incomplete atoms, the ex-post verifier receives the wrong structure.

Latency and Explanation Trade-offs Table 1 also shows the runtime trade-off. Sentence-transformer and cross-encoder baselines are much faster, all below 0.3 seconds on average; the strongest neural baseline, nli-MiniLM2-L6-H768, is also the fastest at 0.130 seconds. LaSSI averages 6.309 seconds, slower than the encoders and llama3.2:3b (3.206 s), roughly level with qwen2.5:7b (6.314 s), but quicker than gemma4:e2b (17.625 s). The additional cost over the encoders is therefore a trade-off for inspectable intermediate structures and directional logical comparison. All timings were measured in the environment described in Section 5; as noted there, the LaSSI figure includes parsing and ontology grounding, so the paradigms are not measured over identical workloads.

8 Limitations and Future Work

At present, the biggest limitation is scale: with 30 evidence items and 90 evidence-claim pairs, one percentage point of overall accuracy corresponds to roughly 0.9 pairs, so the 2.3-point gap between LaSSI and the strongest cross-encoder amounts to two additional correct predictions, and the per-strategy slices in Table 2 rest on as few as $n=4$ items. The reported rankings should therefore be read as indicative of qualitative behavioural differences between paradigms; we report slice sizes alongside every percentage for this reason.

Similarly, the benchmark is deliberately controlled: every evidence item carries exactly one Supported, one Refuted, and one Not Enough Evidence claim, each authored by the first author (and

independently checked by the second) through a predefined modification strategy. This design isolates individual reasoning phenomena and guarantees balanced labels, which is what makes the diagnostic slices interpretable, but the resulting claim distribution is not representative of naturally occurring civic claims, which are noisier and unevenly distributed. **CURB** therefore presents a controlled probe of fine-grained verification behaviour; collecting naturally occurring claims is an explicit goal of the planned larger release. The asymmetric support resolution of Section 5 is likewise part of the benchmark definition rather than a judgement any system makes alone.

LaSSI's structural-rewrite rules and the **HOnK** lifecycle and paraphrase vocabulary were refined using **CURB**, whereas the neural baselines received no domain adaptation. **LaSSI**'s figures are therefore an *optimistic in-domain development estimate*. This is an approximation of what explicit rule and ontology engineering can achieve for these civic forms. Evaluating the frozen system on held-out civic data is the direct test and is left for the larger release.

LaSSI's logical verification is bound by the quality of the upstream parse and grounding. Fragmented civic text, especially roadworks and short-hand traffic updates, can cause dependency parsers to misidentify heads, subjects, temporal modifiers, or measurement phrases. When the syntactic tree is malformed, the resulting logical atoms may fail to capture the intended meaning, leading to incorrect or partial entailment scores.

The fixed thresholds for sentence transformers, rerankers, and other continuous-score outputs are baseline choices rather than optimised decision boundaries, acting as off-the-shelf diagnostics. The post-hoc oracle analysis in Appendix B bounds how much threshold tuning could ever recover for the bi-encoders (at most 64.4% accuracy, still worse than cross-encoders), but tuned figures for all paradigms on a proper validation split are left as future work because the current benchmark is too small for robust tuning without overfitting.

Roadworks notices expose this limitation most clearly: imperative fragments, elided subjects, abbreviations, colon labels, and dense date/time windows can contain several compressed events whose relationships are flattened or lost in the downstream logical form, even after structural chunking, protected date/time spans, and role-priority merging.

The approach depends on **HOnK** coverage.

Typed location, status, weather, paraphrase, and lifecycle resources make the reasoning inspectable, but missing entries or incorrect matches can cause the verifier to miss contradictions or infer too little. This is a principled limitation of ontology-grounded systems: they are transparent, but they require maintenance as new civic forms and local entities appear.

Telegraphic civic notices, especially roadworks, remain difficult for dependency-parser-led logical rewriting, which motivates future work on domain-specific preprocessing or neuro-symbolic semantic parsing rather than reliance on conventional dependency parses. Future work should include broader civic source coverage, a larger and more geographically diverse benchmark, and a clearer interface between neural structured extraction and symbolic verification, including a systematic study of whether ontology-grounded prompting can transfer **HOnK**'s contradiction signals to **LLMs**. Rule-based information extraction (Liu et al., 2024) and few-shot **LLM** schema extraction are possible alternatives to full dependency parsing, but they introduce their own risks, including reduced coverage, brittle patterns, and hallucinated fields (Tenckhoff et al., 2026).

9 Conclusion

This paper introduces **CURB**, a benchmark for fine-grained civic evidence-to-claim reasoning. It is designed to test whether systems can distinguish strict support from contradiction or insufficient evidence when claims differ by local entities, times, causes, statuses, or unsupported details.

The results show that **LaSSI**, supported by **HOnK** grounding, is strongest where the task depends on inspectable directional structure: location mismatch, time mismatch, and cause mismatch. **LaSSI** reaches 85.6% accuracy and 0.858 Macro-F1, narrowly ahead of strong **NLI** cross-encoder baselines (up to 83.3%). The analysis also identifies clear limits: neural models remain competitive on paraphrase and stronger on status changes, and telegraphic roadworks notices remain difficult for the symbolic pipeline, to the point that a cross-encoder overtakes **LaSSI** on that domain. The contribution is therefore not a claim of universal symbolic superiority, but evidence that explicit, logic-grounded structure can reduce false support judgements in civic cases where small local details determine the label.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/W524700/1]. The authors also acknowledge the EPSRC National Edge AI Hub for Real Data: Edge Intelligence for Cyber-disturbances and Data Quality at Newcastle University [grant number EP/Y028813/1].

This research was made possible through the use of open civic data and public information feeds. The authors acknowledge Nexus Tyne and Wear, the Met Office, Newcastle City Council, and Police.uk as the retained public data sources used in the creation of the CURB benchmark via the NEET dataset pipeline.

We also thank the creators and maintainers of ConceptNet, WordNet, Wiktionary, and GeoNames for making these knowledge resources publicly available for use in HOnK.

We further acknowledge Hugging Face and OpenAI for providing access to the publicly available sentence transformer models and LLMs used in the evaluation.

References

- Daniele Battista, Francesco Colace, Angelo Lorusso, Domenico Santaniello, Allan Tucker, and Carmine Valentino. 2025. [Scalable strategy for fake news detection in smart cities combining ai and semantic web technologies](#). In *Proceedings of Tenth International Congress on Information and Communication Technology*, pages 427–437, Singapore. Springer Nature Singapore.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). *Conference on Empirical Methods in Natural Language Processing*, arXiv:1508.05326.
- Swarat Chaudhuri, Kevin Ellis, Oleksandr Polozov, Rishabh Singh, Armando Solar-Lezama, and Yisong Yue. 2021. [Neurosymbolic programming](#). *Foundations and Trends in Programming Languages*, 7(3):158–243.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *Annual Meeting of the Association for Computational Linguistics*, arXiv:2402.03216.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. [The pascal recognising textual entailment challenge](#). In *Machine Learning Challenges*, MLCW’05, page 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Stefano De Giorgis, Aldo Gangemi, and Alessandro Russo. 2025. [Neurosymbolic graph enrichment for grounded world models](#). *Information Processing & Management*, 62(4):104127.
- Wikimedia Foundation. 2026. Wiktionary. <https://www.wiktionary.org/>. Accessed: 29th May 2026.
- Oliver Robert Fox, Giacomo Bergami, and Graham Morgan. 2025. [Verified language processing with hybrid explainability](#). *Electronics*, 14(17).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *arXiv*, arXiv:2407.21783.
- Zhiqiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. 1997. [Fastus: A cascaded finite-state transducer for extracting information from natural-language text](#). *Finite-State Language Processing*, arXiv:cmp-lg/9705013.
- HuggingFace. 2025a. Huggingface - nli-deberta-v3-base. <https://huggingface.co/cross-encoder/nli-deberta-v3-base>. Accessed: 8th June 2026.
- HuggingFace. 2025b. Huggingface - nli-minilm2-l6-h768. <https://huggingface.co/cross-encoder/nli-MiniLM2-L6-H768>. Accessed: 8th June 2026.
- Daniel Z. Korman, Eric Mack, Jacob Jett, and Allen H. Renear. 2018. [Defining textual entailment](#). *Journal of the Association for Information Science and Technology*, 69(6):763–772.
- Ernest Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. 2025. [Knowledge graphs, large language models, and hallucinations: An nlp perspective](#). *Journal of Web Semantics*, 85:100844.
- Xiang Li, Jeffrey Wang, and Lihua Zhang. 2025. [Named entity recognition for smart city data streams: Enhancing visualization and interaction](#). In *International Conference Civil Engineering and Architecture*, volume 13, page 508. Madison Academic Press.
- Pai Liu, Wenyang Gao, Wen Dong, Lin Ai, Songfang Huang, and Yue Zhang. 2024. [A survey on open information extraction from rule-based model to large](#)

- language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9586–9608. Association for Computational Linguistics.
- Md Motaleb Hossen Manik and Ge Wang. 2026. [Unified deployment-aware evaluation of open reasoning language models](#). *arXiv*, arXiv:2604.07035.
- George A. Miller. 1995. [Wordnet: a lexical database for english](#). *Commun. ACM*, 38(11).
- Ollama. 2026. Ollama. <https://ollama.com/>. Accessed: 8th June 2026.
- Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2019. [Universal dependency parsing from scratch](#). In *Conference on Computational Natural Language Learning*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). *Conference on Empirical Methods in Natural Language Processing*, arXiv:1908.10084.
- Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from Wikipedia. In *International Conference on Language Resources and Evaluation*.
- G. Savova, James J. Masanz, P. Ogren, Jiaping Zheng, S. Sohn, Karin Kipper Schuler, and C. Chute. 2010. [Mayo clinical text analysis and knowledge extraction system \(ctakes\): architecture, component evaluation and applications](#). *Journal of the American Medical Informatics Association*, 17(5):507–513.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [Mpnet: Masked and permuted pre-training for language understanding](#). *Neural Information Processing Systems*, arXiv:2004.09297.
- Peiyang Song, Pengrui Han, and Noah Goodman. 2026. [Large language model reasoning failures](#). *Trans. Mach. Learn. Res.*, arXiv:2602.06176.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). *AAAI Conference on Artificial Intelligence*, 31(1).
- Sönke Tenckhoff, Mario Koddembrock, and Erik Rodner. 2026. [Llmstructbench: Benchmarking large language model structured data extraction](#). *arXiv.org*, arXiv:2602.14743.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). *Neural Information Processing Systems*, arXiv:2002.10957.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *arXiv.org*, arXiv:2412.15115.
- Mike Zhang, Johannes Bjerva, and Russa Biswas. 2025. [Follow the path: Reasoning over knowledge graph paths to improve large language model factuality](#). *Annual Meeting of the Association for Computational Linguistics*, arXiv:2505.11140.

A Pipeline Resources and Procedures

This appendix documents the curated resources that the **LaSSI/HONK** pipeline relies on. We include them for transparency: the pipeline’s robustness on civic notices comes from these finite, inspectable rule and vocabulary sets, which also bound how far the current system can be expected to generalise beyond the civic forms in **CURB**.

A.1 Dataset Schema

Table A1 gives the per-row schema of the released **CURB** dataset. The released CSV distinguishes eight `claim_type` values; for reporting, `direct_support` and `approximate_paraphrase` are merged into the single *Paraphrase* strategy of Table 2 ($13 + 17 = 30$ pairs), and `outcome_mismatch` is reported as *Status Change*.

Table A1: Dataset CSV schema.

Column	Description
<code>item_id</code>	Unique identifier including domain and intended label suffix.
<code>domain</code>	Civic category: transport, roadworks, weather, or crime.
<code>evidence_text</code>	Raw factual snippet from the source.
<code>source_name</code>	Source name.
<code>source_page</code>	Page or feed within the source from which the snippet was scraped.
<code>source_url</code>	Source URL or endpoint used for verification.
<code>date_time</code>	ISO 8601 scrape timestamp.
<code>location</code>	Extracted geographic focal point where available.
<code>claim_text_label</code>	Manually authored claim text. Gold label: <i>Supported</i> , <i>Refuted</i> , or <i>Not Enough Evidence</i> .
<code>claim_type</code>	Claim modification strategy.

A.2 Claim Modification Strategies

Each evidence item yields three claims, one per label, produced by the modification strategies in Table A2.

Table A2: Claim generation and modification strategies.

Modification Strategy	Description
Paraphrase	Restates the evidence while preserving the same material truth conditions.
Negation	Introduces a negation that contradicts the core assertion.
Time Change	Alters a temporal bound, timestamp, date range, or expected end date.
Location Change	Substitutes a local entity or changes spatial granularity in a way that affects support.
Cause Change	Changes the stated reason for an event or disruption.
Status/Outcome Change	Changes a lifecycle or service-status value, such as open/closed, delayed/normal, or under-investigation/charged.
Unsupported Extra Detail	Adds plausible but unverified information to an otherwise compatible claim.

A.3 Structured Pre-Parser

Before dependency parsing, a row-chunk pre-parser adapts the terse, semi-structured civic notices for the parser. It first protects atomic spans: ISO timestamps, date ranges, percentages, decimals, and unit-bearing measurements. They survive tokenisation intact, then segments each row and profiles the resulting chunks into roles (for example report header, attribute, status, time range, or action). Verbless metric bodies are rejoined onto their labels, imperative action runs are split, and the chunk-level kernels are merged back into a single representation per row. The stage also normalises scraping artefacts: the crime feed, for instance, emits phrases such as “*on On Or Near Edward Place*”, where consecutive duplicate closed-class words are collapsed and the chained preposition “on or near” is lifted into a single compound spatial marker. The later logical phase then assigns this a disjunctive *at/near* spatial type, which is what preserves the asymmetric spatial entailment discussed in Section 7.

A.4 Structural-Rewrite Rule Inventory

The pipeline applies an ordered registry of 30 structural-rewrite rules that repair the (inherently noisy) dependency-tree-derived kernels into clean logical forms. Of these, 14 are declarative *premise* → *consequence* rules and 16 are bespoke routines reserved for rewrites that need heavier graph surgery. Rather than enumerate internal rule names, we describe the six families of repair they perform; together they cover the civic forms observed in CURB, and coverage of forms outside this set is future work.

Copula, existential and participial recovery re-

builds the main predicate when the parser has mishandled a copula (“is”/“are”), an existential (“there is”), or a participial phrase. For example, reconstructing “*the footbridge is out of use*” when the state phrase was left stranded from its subject. *Relative-clause, conjunction and target restructuring* reassembles coordinated actions (“abandon ... and replace ...”), lifts reduced relative clauses, reattaches stray modifiers, and flattens redundant nesting so that each event has a clean subject and object. *Spatial folding* collects scattered location mentions into a single typed place slot, for instance folding a trailing “in Newcastle” or a geographic modifier into the event’s location. *Causal and temporal canonicalisation* reconstructs cause–effect links (“... resulted in ...”), canonicalises dates and times, and removes duplicated temporal expressions. *Lifecycle and status promotion* attaches a status phrase such as “under investigation” or “out of use” to the correct predicate so that the contradiction detection of Appendix A.5 can fire. Finally, *quantity, weather and measurement* normalisation merges compound quantities and weather-condition modifiers, such as “27% chance of rain” or “light rain shower”, into single typed values.

A.5 Lifecycle Dimensions and Partitions

Categorical status reasoning (Section 7) is implemented as a dimension–partition ontology: each status phrase belongs to a dimension and to a partition within it, and two phrases contradict only when they occupy the same dimension under different non-descriptive partitions (a *descriptive* partition, e.g. “recorded”, never triggers a contradiction). Table A3 lists the 7 dimensions and example phrases for their mutually exclusive partitions (63 phrases in total).

Table A3: Lifecycle dimensions and example phrases per partition. Phrases drawn from different non-descriptive partitions of the same dimension are treated as contradictory.

Dimension	Mutually exclusive partitions (example phrases)
CaseLifecycle	<i>open</i> : under investigation, ongoing <i>closed</i> : charged, concluded, dismissed
SuspectResolution	<i>present</i> : suspect, arrested <i>absent</i> : no suspect identified, unidentified
TransportAvailability	<i>available</i> : operational, in use <i>unavailable</i> : out of use, out of service
WeatherPrecipitation	<i>wet</i> : rain, precipitation <i>dry</i> : dry, arid
SkyCondition	<i>overcast</i> : cloudy, overcast <i>clear</i> : clear skies, sunny
PrecipitationIntensity	<i>high</i> : heavy, torrential <i>low</i> : very low
ParkingCapacity	<i>reduced</i> : reduced, reduction <i>normal</i> : operating normally

A.6 Extended-FOL Construct Routing

Prepositions, markers, and clausal contexts are routed into typed logical constructs, realised as 53 attachment specifications (26 attaching at the event/kernel level and 27 at the argument/Singleton level), plus 54 derivation rules. The subset of constructs that distinguish two events for the ex-post comparison is *derived*, not hand-listed: it is the kernel-attaching constructs, minus a *circumstantial* set that is compared point-wise or paraphrastically (space, time, temporal_context, aim_objective, time_status), plus an *identity-refining* set that distinguishes symmetrically (specification, togetherness, disruption).

A.7 Ex-Post Directional Comparison

For two sentences represented as logical formulae S_i and S_j , the directional support value from S_i to S_j is computed by [Binary-Decision-Diagram \(BDD\)](#) model counting over the minimal atomic constituents, then adjusted by a distinguishing-content cap. Algorithm 1 gives the procedure.

A.8 Worked Examples: A Success and a Failure

We contrast one case the pipeline handles correctly (Figure A1) with one it does not (Figure A2). Each is shown as its directional support matrix, where row 0 is the evidence and rows 1–3 are the Supported, Refuted, and Not Enough Evidence claims;

Table B1: Post-hoc oracle threshold sensitivity for the bi-encoder baselines: accuracy (%) and Macro-F1 under the fixed diagnostic thresholds (0.2, 0.8) versus the best values achievable by any $(\tau_{\text{low}}, \tau_{\text{high}})$ pair selected *with full knowledge of the test labels* ($n = 90$). Oracle figures are an unattainable upper bound.

Model	Default		Oracle best	
	Acc.	F1	Acc.	F1
all-MiniLM-L6-v2	45.6	0.369	64.4	0.647
all-MiniLM-L12-v2	44.4	0.362	53.3	0.540
all-mpnet-base-v2	47.8	0.382	52.2	0.478

reading the evidence row gives the predicted label for each claim.

B Post-Hoc Oracle Threshold Sensitivity

This appendix is a *post-hoc oracle sensitivity analysis*: the thresholds examined here are selected with full knowledge of the test labels, so the resulting figures are an upper bound that no deployable classifier over the same scores could legitimately attain. Its purpose is to answer whether the low bi-encoder accuracy stems from the embeddings, or merely from the particular thresholds chosen, raised by the fixed cut-offs of Section 5.

The bi-encoder baselines map a symmetric cosine similarity to the three [CURB](#) labels through fixed diagnostic thresholds (scores $\leq \tau_{\text{low}} \rightarrow \text{Refuted}$, scores $\geq \tau_{\text{high}} \rightarrow \text{Supported}$, otherwise *Not Enough Evidence*). We re-threshold the cached per-pair similarity scores over the full grid of $(\tau_{\text{low}}, \tau_{\text{high}})$ pairs with $0 \leq \tau_{\text{low}} < \tau_{\text{high}} \leq 1$ in steps of 0.05, and report the best accuracy any pair achieves on the test set itself.

Table B1 shows that better thresholds do help, the best oracle setting lifts all-MiniLM-L6-v2 from 45.6% to 64.4%, but even this unattainable bound remains 19 points below the strongest cross-encoder (83.3%) and 21 points below [LaSSI](#) (85.6%), and the other two bi-encoders peak in the low fifties. The optimal cut-offs are also revealing: for all three models the best pairs sit high and close together (e.g. $\tau_{\text{low}} = 0.75$, $\tau_{\text{high}} = 0.85$ for both MiniLM variants), meaning nearly the entire similarity range must be mapped to *Refuted* because the cosine scores of supported, refuted, and unverifiable claims are compressed into a narrow band near the top of the scale. The deficit is therefore representational — a single symmetric similarity score does not separate the three directional verdicts — rather than an artefact of the particular thresholds chosen in Section 5.

Algorithm 1 Ex-post directional comparison. k_i, k_j are the distinguishing logical-context keys of S_i, S_j (Appendix A.6); rel is the surface relation. The result maps $1.0 \rightarrow \text{Supported}$, $0.0 \rightarrow \text{Refuted}$, and intermediate values $\rightarrow \text{Not Enough Evidence}$.

```

1: function SCORE( $i, j$ )
2:   if hard SPACE contradiction( $i, j$ ) then
3:     return 0.0
4:   end if
5:    $C \leftarrow \top$  ▷ conjunction of pairwise constraints
6:   for all  $a \in \text{atoms}(S_i), b \in \text{atoms}(S_j), a \neq b$  do
7:      $v \leftarrow \text{COMPARE}(a, b)$  ▷ via HOnK/TBox
8:     if  $v = \text{Implying}$  then
9:        $C \leftarrow C \wedge (\neg a \vee b)$ 
10:    else if  $v = \text{Conflicting}$  then
11:       $C \leftarrow C \wedge (a \oplus b)$ 
12:    else if  $v = \text{Equivalent}$  then
13:       $C \leftarrow C \wedge \neg(a \oplus b)$ 
14:    end if
15:  end for
16:   $\text{val} \leftarrow \text{count}(S_i \wedge C \wedge S_j) / \text{count}(S_i \wedge C)$ 
17:  if  $(k_j \setminus k_i) \neq \emptyset$  or identity-refining keys differ or  $(\text{rel}_i \neq \text{rel}_j \wedge k_i \neq k_j)$  then
18:     $\text{rev} \leftarrow \text{SCORE}(j, i)$ 
19:     $\text{val} \leftarrow \min(\text{val}, \text{rev})$  ▷ cap
20:    if  $\text{val} \geq 1.0$  then
21:       $\text{val} \leftarrow 0.5$ 
22:    end if
23:  end if
24:  if  $\text{val} > 0.5 \wedge \text{SPACE relation narrows}$  then
25:     $\text{val} \leftarrow 0.5$ 
26:  end if
27:  return  $\text{val}$ 
28: end function

```

C Confusion Matrices

The label-level confusion matrices for the representative models are given in Figure C1.

all-mpnet-base-v2				nli-deberta-v3-base			
	Sup.	Ref.	NEE		Sup.	Ref.	NEE
Sup.	26	0	4	Sup.	19	0	11
Ref.	6	0	24	Ref.	3	25	2
NEE	13	0	17	NEE	1	0	29

gemma4:e2b				LaSSI			
	Sup.	Ref.	NEE		Sup.	Ref.	NEE
Sup.	27	1	2	Sup.	22	0	8
Ref.	1	26	3	Ref.	0	26	4
NEE	3	16	11	NEE	0	1	29

Figure C1: Confusion matrices (rows = gold, columns = predicted) for representative models on the CURB test split. Sup. = Supported, Ref. = Refuted, NEE = Not Enough Evidence.

Confusion Matrix: transport_005 (Logical)

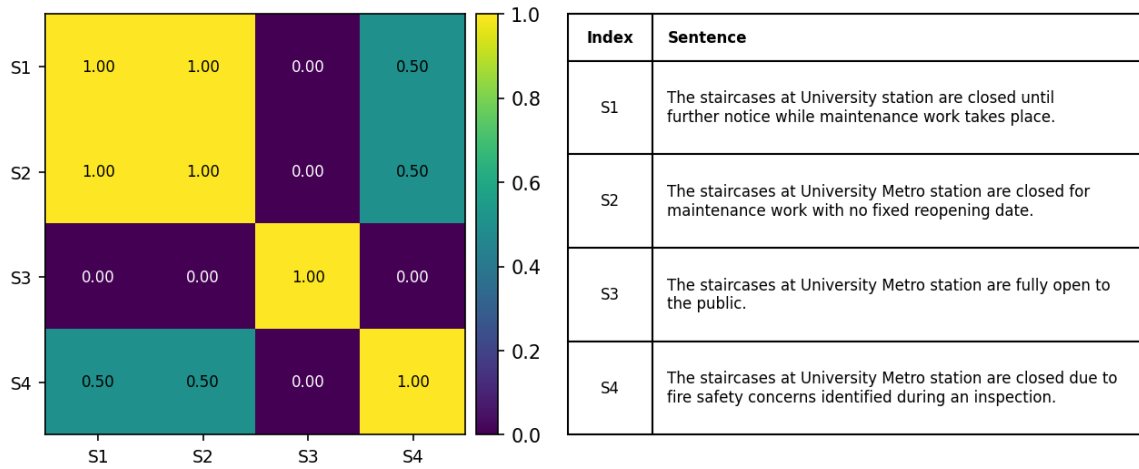


Figure A1: transport_005 (success): the evidence row separates the Supported (1.0), Refuted (0.0), and Not Enough Evidence (0.5) claims correctly. The evidence is “*University staircase work ... The staircases at University station are closed until further notice while maintenance work takes place.*”. All three claims are scored correctly (Figure A1). The supported paraphrase (“*closed for maintenance work with no fixed reopening date*”) scores 1.0; the negation (“*fully open to the public*”) is a hard 0.0, because *open* and *closed* fall in different partitions of the transport-availability lifecycle dimension (Appendix A.5); and a claim adding an unstated cause (“*due to fire safety concerns*”) scores a partial 0.5 and is correctly labelled *Not Enough Evidence*. The case is well behaved because the leading place-name header (“*University staircase work*”) is split from a grammatical clause, leaving a clean predicate for the later phases.

Confusion Matrix: roadworks_002 (Logical)

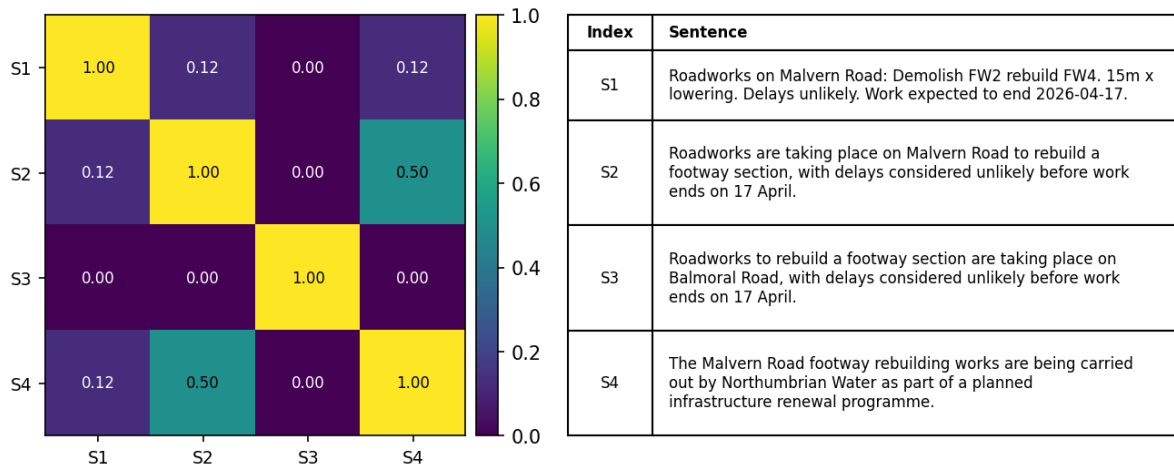


Figure A2: roadworks_002 (failure): the supported claim under-separates to 0.125 (mislabelled *Not Enough Evidence*), while the location-mismatch claim is still correctly refuted at 0.0. The evidence is “*Roadworks on Malvern Road: Demolish FW2 rebuild FW4. 15m x lowering. Delays unlikely. Work expected to end 2026-04-17.*” Here the genuinely supported paraphrase (“*Roadworks ... on Malvern Road to rebuild a footway section*”) scores only 0.125 and is mislabelled *Not Enough Evidence* (Figure A2). The cause is the telegraphic action sequence: bare juxtaposed imperatives (“*Demolish FW2 rebuild FW4*”) and dimensioned abbreviations (“*15m x lowering*”) do not reassemble into clean action atoms, so the supported paraphrase cannot be matched. Notably, the location-mismatch claim (“*... on Balmoral Road*”) is still a hard 0.0 contradiction: the spatial comparison stays robust even when the action structure is lost, so the failure is one of action reassembly, not spatial reasoning.